Theory of Optimal Reshaping with Application to Data Mining

氏名（本籍地）	CHUN JINHEE 全 眞嬉
学位の種類	博 士（情報科学）
学位記番号	情 博 第 368 号
学位授与年月日	平成18年9月7日
学位授与の要件	学位規則第4条第1項該当
研究科、専攻	東北大学大学院情報科学研究科（博士課程）システム情報科学専攻
学位論文題目	Theory of Optimal Reshaping with Application to Data Mining （最適整形の理論とデータマイニングへの応用）
論文審査委員	（主査）東北大学教授 徳山 豪 東北大学教授 篠原 歩 東北大学教授 西関 隆夫 韓国KAIST教授 Kyung-Yong Chwa

論 文 内 容 の 要 旨

1 Introduction

This thesis deals with geometric optimal reshaping problems with applications to data mining. The geometric reshaping problem is a basic problem of computational geometry. However, in the computational geometry approach, the optimization criteria are given according to traditional geometric applications (e.g. pattern matching), therefore they are not suitable for data analysis.

We present a new optimization problem of reshaping input geometric curve representing data distribution considered in data mining application. Namely, we formulate the problem into a problem of peak reducing by reshaping a curve/surface, and design efficient algorithms based on computational geometry and graph algorithm theory.

In our formulation of the optimization problem, we have a family $\mathcal{F}$ of well-shaped geometric objects, and given an input geometric object f, we find an object $\phi \in \mathcal{F}$ that approximates f best. The problem depends on $\mathcal{F}$ and the quality measurement of approximation.

We assume that objects are represented by the trajectories of functions, and we seek for the ϕ that minimizes the distance between f and ϕ with respect to a functional distance such as the L_p-distance. We focus on the number of maximal peaks of the functions/objects to define $\mathcal{F}$. We call *pyramid approximation* if $\mathcal{F}$ is the family of unimodal functions, and *k-peak approximation* if $\mathcal{F}$ is the family of functions that have at most k maximal peaks.

We give a linear time algorithm to find the optimal pyramid approximation of an input piecewise linear function f in one variable under the L_2 distance, and an $O(n \log n)$ time algorithm under the L_p-distance. We also give algorithms to approximate f with a function consisting of the minimum number of unimodal pieces such that the distance is bounded by a given threshold.

For the high-dimensional (i.e. multivariate) problem, we consider the pyramid approximation problem of approximating a function on a d-dimensional voxel grid to minimize the weighted L_2-distance with respect to a given measure distribution on the grid.

Our algorithms are based on computational geometric tools such as convex hull trees as well as graph algorithms.

Finally, we propose a novel construction method for decision trees using the pyramid approximation. This new method can be used for solving the problem of overfitting at the time of constitution a decision tree. When the optimal decision tree is used to training data, prediction accuracy trade-off over unknown data will fall will arise. For reducing the trade-off, we employ the *expert-guided decision tree*. We use the output pyramid

approximation as layered-structure numeric association rules that work as experts of the decision tree. We give theoretical analysis of error bound of our proposed system, as well as experimental report on real data sets.

2 Preliminaries

In this chapter, we introduc computational geometry, graph theory data base, data mining and decision tree.

We consider geometric reshaping problems. The *convex hull* construction problem is a classic problem in computational geometry. A convex hull $C(P)$ of an geometric object P is the smallest convex object containing it. We often identify the convex hull with its boundary curve $C(P)$. We adopt the popular convention to call $C(P)$ the *convex hull* in this thesis. The boundary curve $C(P)$ is represented as a chain (in precise, linked-list structure) of edges and vertices. If we cut $C(P)$ at its leftmost point and the rightmost point, we obtain two chains, one is convex and the other is concave. The lower-located chain (the convex one) is called the *lower hull* of P.

We use grid regions in order to formulate our multivariate function reshaping problems. A function with two variables defined on $[0,1] \times [0,1]$ is called a *grid function* if it takes a constant value in each pixel. Similarly, the partition of the unit d-dimensional cube $[0,1]^d$ into n^d (d-dimensional) squares of side length $1/n$ is called the *voxel grid*, and each small square is called a *voxel*. We use a convention to denote N for the grid size n^d.

We briefly explain graph-theoretic concepts employed in this thesis. Given an undirected graph $G = (V, E)$ and two nodes s and t, an *s-t cut* of G is the separation of V into a pair X and $Y = V \setminus X$ of sets of vertices such that $X \ni s$ and $Y \ni t$. Each edge between a vertex of X and a vertex of Y is called a *cut edge*. If the graph has an edge weight function w', the weight of the cut is the summation of edge weights of cut edges. The minimum *s-t* cut is the *s-t* cut with the minimum weight.

The problem of computing the minimum *s-t* cut is a famous problem in combinatorial optimization and it is also well-known that the computation of the maximum domination closure in a directed graph is reduced to the problem of computing the minimum *s-t* cut of a graph obtained from the directed graph.

3 Our motivation and problems

A *data record* in a database can be considered as a vector (*data tuple*) of numerics and symbols. If we give some key attribute values (such as ID or name), we can obtain all the attribute values of the data record by using database query. Hence, database can work as functions. Moreover, if we give a vector of values of a set of attributes as a query input, we can report all data satisfying the query condition. Therefore, database management system is used to organize the database that it can efficiently work when a user wants to use it as a function. Indeed, there are many implicit functions represented in a database and we do not even know which one is useful for users. Data warehousing is a system that can handle this problem.

Data mining aims to extract knowledge from a database. In a sense, it is a reshaping problem of the database into a set of rules (typically, association rules) for decision systems. Here, both the association rules and decision systems are functions. For example, if we have a conventional association rule $C_1 \to C_2$ and a new data t, then we report $C_2 = yes$ if t satisfies C_1, otherwise $C_2 = no$. This is a very simple function that take values 0 and 1. Therefore, the data mining problem can be considered a (very sophisticated) variant of reshaping problem.

As we discuss above, data mining is neither a reshaping problem of a single function nor a geometric problem. Moreover, the problem description itself is obscure. Nevertheless, we need a tool to reform data distribution information into a well-behaved function and we would like to demonstrate that function reshaping can be utilized as a tool for designing a data mining system. Along this philosophy, as the second main topic, we propose the *layered rule* that is in the same form $C_1 \to C_2$ as the conventional rule, but given a new data t it returns a value $f(t)$ to represent the estimated confidence value. The optimal layered rule is defined and computed based on ideas of function reshaping and the rule is applied to constructing a novel expert-guided decision system.

4 One dimensional reshaping problem in L_2 metric

In this chapter, we discuss peak-reducing fitting problem of a curve under the L_2 metric.

Given a function $y = f(x)$ in one variable, we first consider the problem of computing the single-peaked (*unimodal*) curve $y = \phi(x)$ minimizing the L_2-distance between them. When the input function f is a histogram with $O(n)$ steps or a piecewise linear function with $O(n)$ linear pieces, we design algorithms for computing

ϕ in linear time. We then give an algorithm to approximate f with a function consisting of the minimum number of unimodal pieces under the condition that each unimodal piece is within a fixed L_2-distance from the corresponding portion of f.

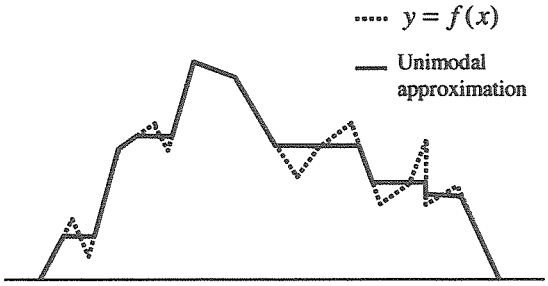

Figure 1: Input polygonal function (dotted curve) and its optimal unimodal approximation.

Given a function $y = f(x)$ defined on an interval $[0, 1]$, we consider the problem of approximating f by a *unimodal function* $y = \phi(x)$. Here, a function is *unimodal* if it has a unique maximal peak (the peak may be a flat interval). Equivalently, for any real number t, $\{x \in [0, 1] : \phi(x) \geq t\}$ is either an interval or empty. We assume that functions considered in this chapter are bounded and *Riemann integrable* (e.g. a piecewise algebraic function). For two functions $g(x)$ and $h(x)$ defined on $[0, 1]$, we define their *inner product* as $g \cdot h = \int_0^1 g(x)h(x)dx$. The *$L_2$-norm* of a function g is $||g|| = \sqrt{g \cdot g}$, and the *L_2-distance* between two functions g and h is $||g - h||$.

We consider the squared L_2-distance

$$||f - \phi||^2 = \int_0^1 (f(x) - \phi(x))^2 dx$$

between f and ϕ. When the squared L_2-distance is minimized, we call ϕ *the optimal unimodal approximation of* f. See Figure 1 to get intuition. We assume without loss of generality that f and ϕ are nonnegative functions, since we can vertically translate them without changing the distance between them.

Theorem 4.1 *Suppose that the optimal unimodal approximation ϕ of f attains its maximum value at $x = p$. Then, the curve defined by $y = \Phi(x)$ coincides with $\mathcal{L}(p)$ and $\mathcal{U}(p)$ in the ranges $[0, p]$ and $[p, 1]$ of x, respectively.*

Theorem 4.2 *The optimal unimodal approximation of a histogram f with n steps can be computed in $O(n)$ time.*

Theorem 4.3 *The optimal unimodal approximation ϕ of a piecewise linear function f with n linear pieces can be computed in $O(n)$ time.*

Theorem 4.4 *A piecewise unimodal approximation with the minimum number of pieces can be computed in $O(n \log n)$ time.*

5 One dimensional reshaping problem in L_p metric

In this chapter, we consider the problem peak-reducing fitting of a curve under the L_p metric.

Given a function $y = f(x)$ defined on an interval $I = [0, 1]$, we consider the problem of approximating f by a k-peaked function $y = \phi(x)$. Here, a function is k-peaked if the function has at most k maximal peaks (each peak may be a flat interval). If the distance between input function f and output ϕ is minimized, we call ϕ the *optimal k-peaked approximation* of f.

In particular, if $k = 1$, we call it the *optimal unimodal approximation.* See Figure 1 to get intuition. We assume that the input function f is piecewise linear with n linear pieces, and evaluate the quality of approximation by the L_p-distance

$$||f, \phi||_p = (D_p(f, \phi))^{1/p} = [\int_0^1 |f(x) - \phi(x)|^p dx]^{1/p}.$$

For example, the L_1 minimization problem is equivalent to minimization of the area of the region bounded by the curves of input and output functions , since the integral of vertical distance becomes the area.

Theorem 5.1 *The algorithm compute $T = \cup_{1 \le i \le n} C_i$ in $O(n \log n)$ time.*

Theorem 5.2 *The L_p-optimal unimodal approximation of a piecewise linear function f with n linear pieces can be computed in $O(n \log n)$ time.*

Theorem 5.3 *Optimal k-peaked approximation ϕ of f is computed in $O(kh^2 + hn \log n)$ time.*

6 Multi-dimensional reshaping problem

In this chapter, we discuss the problem of approximating a function on a d-dimensional voxel grid by a unimodal function to minimize the L_2 approximation error with respect to a given measure distribution on the grid. The output unimodal function gives a layered structure on the voxel grid, and we give efficient algorithms for computing the optimal approximation under a reasonable assumption on the shape of each horizontal layer. Our main technique is a dominating cut algorithm for a graph.

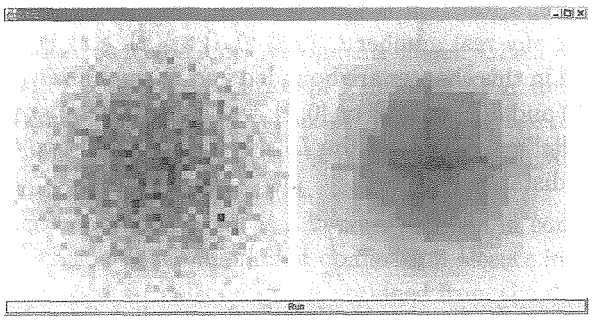

Figure 2: Pyramid approximation (right) of a 2-dimensional terrain (left).

Given a geometric object, transforming it into an object satisfying a certain geometric property is a problem frequently discussed in computational geometry. In this chapter, we consider the following problem: Consider a function f in d variables as an input. We want to transform the function into a unimodal function $\phi(x)$ minimizing the L_2 error (with respect to a given measure distribution) under the condition that each horizontal slice $\{x : \phi(x) \le t\}$ at height t has a good geometric shape. The computational complexity of the problem depends on the representation of input and output functions. For simplicity, we assume that the functions are defined on a voxel grid and investigate the complexity based on the number of voxels, although our framework works if f is defined on other types of subdivisions of $\mathbb{R}^d$ such as triangulations by giving suitable modification.

If $d = 2$, the optimal pyramid can be considered as reshaping a terrain defined by f into a single-peak terrain defined by ϕ by moving earth from higher position to lower position minimizing the loss of positional potential. Figure 2 illustrates pyramid approximation (right) of a 2-dimensional terrain (left), where μ is constant and $f(c)$ and $\phi(c)$ are represented by using the gray levels of pictures. Here, black and white represent 1 and 0, respectively. This seems to be a basic problem in computational geometry and geography.

In this chapter, we investigate the condition under which the optimal pyramid can be efficiently constructed. Our condition is defined by using dominating sets of a directed graph. We can flexibly control the size of the family of regions by using the graph. In particular, we define the following d-dimensional region families: 1) stabbed-union of orthogonal regions, (2) generalized base-monotone regions, and (3) digitized star-shaped regions. Surprisingly, for each of these region families, the optimal pyramid can be computed in polynomial time in the number n of voxels by reducing the problem to the minimum s-t cut problem in a directed graph.

Theorem 6.1 *The optimal pyramid for $\mathcal{R}$ of M different regions can be computed in $O(M^2 n)$ time.*

Theorem 6.2 *The optimal pyramid for the family $\mathcal{R}_H$ can be computed in $O(n^{1.5} \log n \log^2 N)$ time. The time complexity is improved to $O(n \log N)$ if H is a tree.*

Theorem 6.3 *The optimal pyramid with respect to the family of domination closures of $T_0(c)$ can be computed in $O(n \log N)$ time.*

Theorem 6.4 *The optimal pyramid for the family of all connected lower half regions can be computed in* $O(N \log^2 N)$ *time.*

Theorem 6.5 *Suppose that the values of* ρ *and* μ *are quotient numbers of integers less than* Γ. *The optimal pyramid for the family of downstep regions can be computed in* $O(N \log(N\Gamma))$ *time.*

Theorem 6.6 *The optimal pyramid for the family of point-stabbed unions at* p *can be computed in* $O(N \log(N\Gamma))$ *time.*

Theorem 6.7 *For all* $p \in \mathcal{G}$, ϵ*-approximations of pyramids of the point-stabbed union at* p *can be computed in* $O(\epsilon^{-1} N^{1.5} \log N)$ *time.*

7 Construction of expert guided decision tree on numeric database

In this chapter, we propose a novel technique for constructing a decision tree with high prediction accuracy on a numeric database.

We propose a technique for constructing a decision tree with high prediction accuracy. There is a problem of overfitting at the time of the decision tree construction. If an optimal decision tree is used to training data, trade-off that the prediction accuracy over unknown data will fall will arise. To avoid this, we employ an *expert-guided decision tree.* We propose a new method for removing these drawbacks. We use layered-structure numeric association rules as the expert of a decision tree. In the expert-guided decision tree, an expert is attached to each branching node, and gives decision of the pruned tree at the node, together with a function $f(t)$ to give the confidence of the expert's decision depending on the input t. In our proposed system, we use the optimal pyramid corresponding to the given data distribution generating the region or interval rule at the node. See Figure 3 for getting intuition.

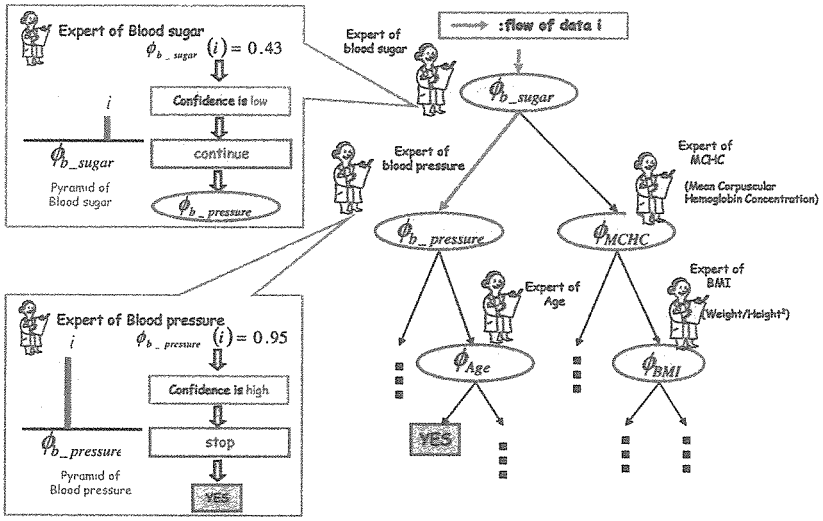

Figure 3: Expert guided decision tree with optimal layered rules

We presents theoretical analysis of error bound of our proposed system, as well as experimental report on real data sets. The aim of an experiment is to evaluate the accuracy of the expert guided decision tree. In this experiment, the optimal pyramid was built to each node of the decision tree before pruning which C4.5 generated. And the expert was given. The expert of each node gives the reliability depending on each data to test data. If the reliability is in the range which the user specified, a label gives at the time. Then this classification is finished and it progresses to the next classification.

8 Conclusions

In this thesis, we have dealt with geometric optimal reshaping problem with application to data mining. We have presented a new optimization problem of reshaping of input geometric curve representing a data distribution considered in data mining application.

論文審査結果の要旨

本論文ではデータマイニングへ応用することを目的として、幾何学図形の整形に関する新しい最適化問題を提案し、そのアルゴリズムを設計し、理論的解析を行うとともに、具体的な応用を論じている。関数の近似理論は古くから数学の大テーマであり、フーリエ解析やウエーブレット解析など数多くの研究がある。一方、理論計算機科学においても図形の整形は計算幾何学（幾何学データ処理）では非常に重要な問題であり、出力図形の複雑度（例えば区分線形関数の節点の個数）の最小化問題などが広く研究されている。著者は、データマイニングで重要なテーマである、数値データベースからのルール抽出の最適化に利用するため、関数として与えられた図形の整形に関して新しい最適化問題を提案し、計算幾何学及びグラフアルゴリズムの手法を用いてその問題を解く理論を構築した。　本論文はこれらの成果を取りまとめたものであり、全編７章からなる。

第１章は序論である。

第２章では本研究で用いる関連事項を解説している。

第３章ではデータマイニングで広く用いられる結合ルール生成における現状の問題点を考察し、それを解決するために新しい図形整形問題を定式化している。具体的には入力関数を単峰関数で最適近似する問題(ピラミッド近似問題)、および極大点がｋ個以下の関数で最適近似する問題(多峰近似問題)を数学的に定式化している。これは、幾何学的最適化における新しい問題である。

第４章では１変数関数を考察し、最も標準的な二乗誤差距離を評価基準とし、凸包を用いてピラミッド近似問題を線形時間で解くアルゴリズムを設計し、更に多峰近似問題に対しO(n log n)時間アルゴリズムを与えている。これらの結果は、計算幾何学の成果として高く評価できる。

第５章では第４章で得られたアルゴリズムを一般化し、Ｌｐ距離を評価基準としてピラミッド近似や多峰近似問題を解く高速なアルゴリズムを設計している。

第６章では多変数関数に対するピラミッド近似問題を考察し、その数理モデル化を与えるとともに、多項式時間で解けることを示している。図形の切断面に対する新しい考察と、ピラミッド近似問題をグラフの最大重み支配閉包問題へ帰着するというアイデアは高く評価できる。

第７章ではデータマイニングへの応用を扱い、ピラミッド近似で得られた出力関数を用いたエキスパート付き決定木の構築、オンライン学習理論を用いたパラメタ調整の実装、さらに理論と実験により学習精度の評価及び解析を行っている。

以上要するに本論文は、新しい幾何学図形整形理論を提案するとともに、それを利用して新しいデータマイニング手法を提案し、その有用性を理論的解析および実験により評価したもので、システム情報科学の発展に寄与するところが少なくない。

よって、本論文は博士(情報科学)の学位論文として合格と認める。